Recency, Popularity, and Diversity of Explanations in Knowledge-based Recommendation

Discussion Paper

Giacomo **Balloccu**, Ludovico **Boratto**, Gianni **Fenu and** Mirko **Marras***

Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

Abstract

Modern knowledge-based recommender systems enable the end-to-end generation of textual explanations. These explanations are created from learnt paths between an already experience product and a recommended product in a knowledge graph, for a given user. However, none of the existing studies has investigated the extent to which properties of a single explanation (e.g., the recency of interaction with the already experience product) and of a group of explanations for a recommended list (e.g., the diversity of the explanation types) can influence the perceived explanation quality. In this paper, we summarize our previous work on conceptualizing three novel properties that model the quality of the explanations (linking interaction recency, shared entity popularity, and explanation type diversity) and proposing re-ranking approaches able to optimize for these properties. Experiments on two public data sets showed that our approaches can increase explanation quality according to the proposed properties, while preserving recommendation utility. Source code and data: https://github.com/giacoballoccu/explanation-quality-recsys.

Keywords

Recommender Systems, Explainability, Evaluation, Responsible Recommendation

1. Introduction

Explaining to users *why* certain results have been provided to them has become an essential property of modern Recommender Systems (RSs). Under certain conditions, these explanations are required by regulations like the European General Data Protection Regulation (GDPR) [1]. Moreover, they have been proved to positively impact on businesses and user experience [2]. These explanations are currently made possible by means of traditional recommender systems augmented with external knowledge, modelled in form of knowledge graphs (KGs).

Existing approaches have incorporated the implicit representation of users and products in a knowledge graph as a regularization term in the objective function of a traditional recommender system (regularized-based approaches) [3, 4, 5, 6, 7, 8] or exploited the structure of the knowledge graph to produce paths between users and products during the optimization process (path-based approaches) [9, 10, 11]. These latter paths can be in turn translated into a short text to be provided to the user for explaining why a certain product has been recommended to them.

Path-based approaches often rely on a reinforcement learning (RL) agent conditioned on a

*Corresponding author.
✉ mirko.marras@acm.org (M. Marras)

user in the knowledge graph and trained to navigate to potentially relevant products for that user. The agent then samples paths between the user and the recommended products in the knowledge graph, based on the probability the agent took that path. In the movie domain, an example path between an already watched movie ($movie_1$) and a recommended movie ($movie_2$), shaped in the form of $user_1$ `watched` $movie_1$ `directed` $director_1$ `directed` $movie_2$, can be used to provide the explanation "$movie_2$ is recommended because you watched another movie directed by $director_1$". These paths include a linking interaction ($user_1$ `watched` $movie_1$), a shared entity ($director_1$), a sharing relation (`directed`), and a recommended item ($movie_2$). However, multiple paths between the user and the recommended item exist, and the path used to create the explanation is merely selected based on an inner-functioning probability so far.

In this paper, we summarize our prior work [12] on conceptualizing three novel user-level explanation quality metrics deemed as important for the next generation of explainable RSs, considering recency, popularity and diversity of explanation. We then propose re-ranking approaches that optimize the top-k recommended products and the paths exploited for generating the accompanying explanations for the proposed explanation quality metrics, going beyond mere inner-functioning probability scores. We finally assess the impact of our approaches on recommendation utility and explanation quality, investigating whether any trade-off exists.

2. Methodology

In this section, we introduce the datasets, the pre-processing steps, the evaluation metrics used to assess recommendation utility and explanation quality, and the re-ranking approaches proposed for their joint optimization. Please refer to the original work [12] for detailed information.

Dataset Preparation. Our experiments were run on two real-world datasets, namely Movielens1M and LastFM. MovieLens1M (ML-1M) [13] contains 1,000,209 ratings assigned to 3,226 movies by 6,040 users of the online service MovieLens. This dataset was extended with its DBpedia KG proposed in [14]. On the other hand, LastFM is a dataset composed of 3,955,598 interactions, 47,981 songs and 15,773 users, extracted by [7] from the original LastFM1b [15]. The associated KG was extracted from Freebase by [16]. For both datasets, we discarded the products (and their corresponding interactions) not present as entities in the KG. For each dataset, we performed a temporal user-based training-validation-test split, with the 70% oldest interactions in the training set, the subsequent 10% in the validation set (hyper-parameter tuning), and the last 20% in the test set. Products already seen by the user were not recommended a second time.

Explanation Property Design. We adopted a mixed approach combining literature review and user's studies to explore and conceptualize the space of relevant explanation quality metrics comprehensively. As a result of this first phase, we identified and operationalized three novel explanation quality metrics. These metrics, whose importance for the users was assessed through a questionnaire, are based respectively on the concepts of recency, popularity and diversity applied to the components of a path in the KG (e.g., u_i `listened` $song_1$ `featuring` $artist_1$ `featuring` $song_2$). Linked interaction recency (LIR) monitors how recent the user interaction with the product included in the selected path is, i.e., u_i `listened` $song_1$ in the example path. Shared Entity Popularity (SEP) considers how popular the shared entity is, i.e., $artist_1$ in the example path. Explanation Type Diversity (ETD) counts how many explanations
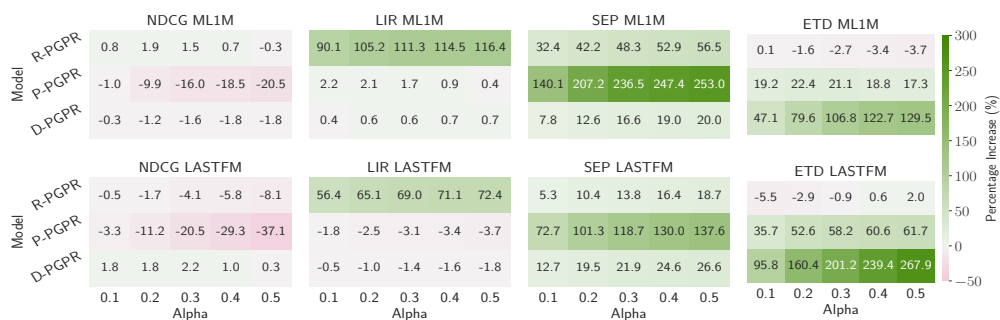
types (e.g., `featured`, `wrote by`, and `composed by` in music) are provided in a top-k list.

Recommendation and Re-Ranking. Being focused on improving the quality of the textual explanations provided to the users, our suite of re-ranking approaches can be applied to the output of any recommender system able to generate paths between users and products in the KG. To this end, in our study, we focused on PGPR [17]. Specifically, our re-ranking approaches were built to re-rank both the recommended products and the explanation paths originally generated by a recommender system, considering both explanation quality scores and embedding scores (as a proxy of recommendation utility). The optimization problem aimed at maximizing the weighted sum of these two factors, where α is an hyper-parameter defining the importance given to explanation quality. Our experiments led to 14 different re-ranked lists and explanations {R|P|D|RP|RD|PD|RPD}-PGPR (7 combinations of optimized explanation quality metrics × two datasets), marked according to the combination of explanation quality metrics the re-ranked recommended list was optimized for (*R*ecency : LIR; *P*opularity : SEP; *D*iversity : ETD).

Evaluation. Our evaluation process covered recommendation utility and explanation quality, computed on top-10 recommended lists ($k = 10$) for the sake of conciseness and clarity. We assessed recommendation utility through the Normalized Discounted Cumulative Gain (NDCG) using binary relevance scores and a base-2 logarithm decay (as done in [18, 19] for example). We assessed explanation quality from three perspectives, namely linking interaction recency (LIR), shared entity popularity (SEP), and explanation type diversity (ETD). For each metric, we reported the average value across the entire user base obtained with the test data.

3. Experimental Results

Trade-off between Recommendation Utility and Explanation Quality (RQ1). First, we investigated the extent to which our re-ranking approaches impacted on explanation quality and recommendation utility, according to the optimized explanation quality metric(s). Figure 1 collects the gains / losses resulting from our re-ranking approaches in terms of NDCG, LIR, SEP, and ETD, over $\alpha \in [0.1, 0.5]$, with respect to the original PGPR model. For conciseness, we focus on only single-metric optimization. Results show that optimizing for a certain explanation quality metric x not only causes gains on x for the resulting explanations, but positively affects

NDCG ML1M

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	0.8	1.9	1.5	0.7	-0.3
P-PGPR	-1.0	-9.9	-16.0	-18.5	-20.5
D-PGPR	-0.3	-1.2	-1.6	-1.8	-1.8

LIR ML1M

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	90.1	105.2	111.3	114.5	116.4
P-PGPR	2.2	2.1	1.7	0.9	0.4
D-PGPR	0.4	0.6	0.6	0.7	0.7

SEP ML1M

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	32.4	42.2	48.3	52.9	56.5
P-PGPR	140.1	207.2	236.5	247.4	253.0
D-PGPR	7.8	12.6	16.6	19.0	20.0

ETD ML1M

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	0.1	-1.6	-2.7	-3.4	-3.7
P-PGPR	19.2	22.4	21.1	18.8	17.3
D-PGPR	47.1	79.6	106.8	122.7	129.5

NDCG LASTFM

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	-0.5	-1.7	-4.1	-5.8	-8.1
P-PGPR	-3.3	-11.2	-20.5	-29.3	-37.1
D-PGPR	1.8	1.8	2.2	1.0	0.3

LIR LASTFM

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	56.4	65.1	69.0	71.1	72.4
P-PGPR	-1.8	-2.5	-3.1	-3.4	-3.7
D-PGPR	-0.5	-1.0	-1.4	-1.6	-1.8

SEP LASTFM

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	5.3	10.4	13.8	16.4	18.7
P-PGPR	72.7	101.3	118.7	130.0	137.6
D-PGPR	12.7	19.5	21.9	24.6	26.6

ETD LASTFM

Model	0.1	0.2	0.3	0.4	0.5
R-PGPR	-5.5	-2.9	-0.9	0.6	2.0
P-PGPR	35.7	52.6	58.2	60.6	61.7
D-PGPR	95.8	160.4	201.2	239.4	267.9

(x-axis: Alpha; y-axis: Percentage Increase (%), range −50 to 300)

Figure 1: Impact of our re-ranking approaches R-PGPR (recency), P-PGPR (popularity), and D-PGPR (diversity) on recommendation utility (NDCG) and explanation quality (LIR, SEP, ETD), w.r.t. PGPR.

Table 1
Recommendation utility (NDCG).

	Baseline ↑			Ours ↑	
Model	ML1M	LASTFM	Model	ML1M	LASTFM
FM	0.32	**0.15**	R-PGPR	**0.34**	<u>0.14</u>
BPR	<u>0.33</u>	0.13	P-PGPR	<u>0.33</u>	0.13
NMF	0.32	<u>0.14</u>	D-PGPR	0.32	<u>0.14</u>
CKE	<u>0.33</u>	<u>0.14</u>	DP-PGPR	0.31	0.13
CFKG	0.27	0.08	PR-PGPR	0.32	0.13
KGAT	<u>0.33</u>	**0.15**	DR-PGPR	0.32	0.13
PGPR	<u>0.33</u>	<u>0.14</u>	DPR-PGPR	0.31	0.13

For each dataset: best result in **bold**, second-best result <u>underlined</u>.

Table 2
Explanation quality (EQ, LIR, SEP, ETD).

	ML1M				LASTFM			
	EQ ↑	LIR ↑	SEP ↑	ETD ↑	EQ ↑	LIR ↑	SEP ↑	ETD ↑
PGPR	0.81	0.43	0.26	0.12	1.07	0.56	0.38	0.13
R-PGPR	1.37	**0.88**	0.37	0.12	1.50	**0.93**	0.43	0.14
P-PGPR	1.22	0.44	0.63	0.15	1.42	0.55	0.67	0.20
D-PGPR	1.05	0.43	0.33	**0.29**	1.50	0.55	0.47	**0.48**
DP-PGPR	1.37	0.44	**0.70**	0.22	1.56	0.55	<u>0.68</u>	0.32
PR-PGPR	**1.71**	<u>0.86</u>	**0.70**	0.14	<u>1.80</u>	0.86	0.50	0.43
DR-PGPR	1.59	0.84	0.46	**0.29**	**2.01**	<u>0.86</u>	**0.69**	<u>0.46</u>
DPR-PGPR	<u>1.70</u>	0.79	<u>0.67</u>	<u>0.23</u>	1.76	0.83	0.63	0.29

For each metric: best result in **bold**, second-best result <u>underlined</u>.

other explanation quality metrics (e.g., optimizing for SEP led to gains on both SEP and ETD and vice-versa). Even with $\alpha =$0.1, our re-ranking approaches led to significant gains ($\geq 50\%$) on the optimized metric, without a significant loss (just $\leq 1\%$) in recommendation utility.

Impact of Producing Explanations on Recommendation Utility (RQ2). We then investigated how the recommendation utility achieved by our re-ranking approaches compared to that achieved by other traditional recommender systems[1]. Table 1 shows the NDCGs obtained by the considered systems. Results show that the recommendations obtained through our re-ranking approaches achieved state-of-the-art NDCGs. In both datasets, all re-ranking approaches achieved a NDCG equal or slightly lower than the non-(path-)explainable baselines. Overall, our results show that accounting for beyond-accuracy aspects on user-level explanation quality does not lead to losses (when observed, they are negligible) in recommendation utility.

Impact on Explanation Quality (RQ3). Under the same setting introduced in the last experiment, we finally investigated how optimizing for a different (set of) explanation quality metrics through our re-ranking approaches impacted on individual and joint explanation quality metrics. Table 2 reports the LIR, SEP, ETD, and EQ (the sum of the three explanation metric scores). Gains in explanation quality were large and proportional to the baseline PGPR value, on both datasets. Higher gains were observed for ETD than other metrics. Considering all three metrics jointly did not lead to the highest overall explanation quality. This highlights possibly diverging optimization patterns according to the domain and characteristics of the dataset.

4. Conclusion

In this paper, we conceptualized and operationalized three novel metrics to monitor explanation quality in recommendation from a user perspective. We then proposed re-ranking approaches to optimize for these properties. Our results on two real-world datasets showed that not only the proposed approaches improve the overall explanation quality, but also preserve (and sometimes improve) recommendation utility w.r.t. the original model. In next steps, we plan to operationalize a larger space of user-level explanation metrics and assess the impact of the resulting explainable RSs on the platform and its stakeholders via online experiments.

---

[1]We selected the α value for our approaches, assuming to work in a context where the platform owners are willing to lose 10% of NDCG at most to increase as much as possible explanation quality

References

[1] B. Goodman, S. R. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (2017) 50–57. URL: https://doi.org/10.1609/aimag.v38i3.2741. doi:10.1609/aimag.v38i3.2741.

[2] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, 2007, pp. 801–810. doi:10.1109/ICDEW.2007.4401070.

[3] F. Zhang, N. J. Yuan, D. Lian, X. Xie, W.-Y. Ma, Collaborative knowledge base embedding for recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 353–362. URL: https://doi.org/10.1145/2939672.2939673. doi:10.1145/2939672.2939673.

[4] H. Wang, F. Zhang, X. Xie, M. Guo, Dkn: Deep knowledge-aware network for news recommendation, 2018. URL: https://arxiv.org/abs/1801.08284. doi:10.48550/ARXIV.1801.08284.

[5] J. Huang, W. X. Zhao, H. Dou, J.-R. Wen, E. Y. Chang, Improving sequential recommendation with knowledge-enhanced memory networks, in: The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 505–514. URL: https://doi.org/10.1145/3209978.3210017. doi:10.1145/3209978.3210017.

[6] G. He, J. Li, W. X. Zhao, P. Liu, J. Wen, Mining implicit entity preference from user-item interaction data for knowledge graph completion via adversarial learning, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 740–751. URL: https://doi.org/10.1145/3366423.3380155. doi:10.1145/3366423.3380155.

[7] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, ACM, New York, NY, USA, 2019, p. 950–958. URL: https://doi.org/10.1145/3292500.3330989. doi:10.1145/3292500.3330989.

[8] Y. Qu, T. Bai, W. Zhang, J. Nie, J. Tang, An end-to-end neighborhood-based interaction model for knowledge-enhanced recommendation, CoRR abs/1908.04032 (2019). URL: http://arxiv.org/abs/1908.04032. arXiv:1908.04032.

[9] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement knowledge graph reasoning for explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 285–294. URL: https://doi.org/10.1145/3331184.3331203. doi:10.1145/3331184.3331203.

[10] W. Song, Z. Duan, Z. Yang, H. Zhu, M. Zhang, J. Tang, Ekar: An explainable method for knowledge aware recommendation, CoRR abs/1906.09506 (2022). URL: http://arxiv.org/abs/1906.09506. arXiv:1906.09506.

[11] K. Zhao, X. Wang, Y. Zhang, L. Zhao, Z. Liu, C. Xing, X. Xie, Leveraging Demonstrations for Reinforcement Recommendation Reasoning over Knowledge Graphs, Association for Computing Machinery, New York, NY, USA, 2020, p. 239–248. URL: https://doi.org/10.1145/3397271.3401171.

[12] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations, CoRR abs/2204.11241 (2022). URL: https://doi.org/10.48550/arXiv.2204.11241. doi:10.48550/arXiv.2204.11241. arXiv:2204.11241.

[13] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015). URL: https://doi.org/10.1145/2827872. doi:10.1145/2827872.

[14] Y. Cao, X. Wang, X. He, Z. Hu, T. Chua, Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences, in: L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, L. Zia (Eds.), The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM, 2019, pp. 151–161. URL: https://doi.org/10.1145/3308558.3313705. doi:10.1145/3308558.3313705.

[15] M. Schedl, The lfm-1b dataset for music retrieval and recommendation, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 103–110. URL: https://doi.org/10.1145/2911996.2912004. doi:10.1145/2911996.2912004.

[16] W. X. Zhao, G. He, K. Yang, H. Dou, J. Huang, S. Ouyang, J.-R. Wen, KB4Rec: A Data Set for Linking Knowledge Bases with Recommender Systems, Data Intelligence 1 (2019) 121–136. URL: https://doi.org/10.1162/dint_a_00008. doi:10.1162/dint_a_00008.

[17] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement knowledge graph reasoning for explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 285–294. URL: https://doi.org/10.1145/3331184.3331203. doi:10.1145/3331184.3331203.

[18] M. Marras, L. Boratto, G. Ramos, G. Fenu, Equality of learning opportunity via individual fairness in personalized recommendations, International Journal of Artificial Intelligence in Education (2021) 1–49. URL: https://doi.org/10.1007/s40593-021-00271-1. doi:10.1007/s40593-021-00271-1.

[19] L. Boratto, G. Fenu, M. Marras, G. Medda, Consumer fairness in recommender systems: Contextualizing definitions and mitigations, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, volume 13185 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 552–566. URL: https://doi.org/10.1007/978-3-030-99736-6_37. doi:10.1007/978-3-030-99736-6_37.